Course Content

Two-sample analysis

Approximations of expectation and variance

Transformations

Regression Analysis

Two-level experiments

Analysis of variance

Statistical Process Control

Chi Square tests for distributions, independence and homogeneity

Non Parametric Statistics

Two sample tests

Random sample based methods (Z-test and t-test) should not be used on correlated data.

Historical data may help but need to be representative.

Approximation of expectation and variance and transformations

Approximation of mean and variance. $Y = f\left(X_1,\ldots X_n\right)$. For independent variables and small variances:

$$E\left[Y\right] \approx f\left(\mu_1,\ldots,\mu_n\right)$$

$$Var\left[Y\right] \approx \sum_{i=1}^{n}\left(\frac{\partial f\left(\mu_1,\ldots\mu_n\right)}{\partial x_i}\right)^2 Var\left[X_i\right]$$

Transformation of variables to obtain constant variance.

$$Y = g\left(X\right) \approx g\left(\mu\right) + g'\left(\mu\right)\left(X-\mu\right)$$

$$Var\left[Y\right] \approx \left(g'\left(\mu\right)\right)^2 Var\left(X\right) = k$$

$$\sigma_X \propto \mu \Rightarrow g\left(X\right) = \ln\left(X\right)$$

$$\sigma_X \propto \mu^2 \Rightarrow g\left(X\right) = \frac{1}{X}$$

$$\sigma_X \propto \sqrt{\mu} \Rightarrow g\left(X\right) = \sqrt{X}$$

Simple linear regression

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \begin{cases} E[\varepsilon_i] = 0,\ \mathrm{Var}[\varepsilon_i] = \sigma^2 \\ \textit{independent} \end{cases}$$

Method of least squares: Minimixe $Q(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$ with respect to b_0 and b_1. This gives:

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad , \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) Y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x} \ , \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

Regression line: $\hat{y} = b_0 - b_1 x$

Residuals: $y_i - \hat{y}_i = y_i - b_0 - b_1 x_i$

Inference in linear regression

$$T_{\beta_1} = \frac{\hat{\beta}_1 - \beta_1}{\frac{S}{\sqrt{S_{xx}}}} \sim t_{n-2}$$

$$T_{\beta_0} = \frac{\hat{\beta}_0 - \beta_0}{S\sqrt{\frac{\sum_{i=1}^{n} x_{1i}^2}{nS_{xx}}}} \sim t_{n-2}$$

where $S_{xx} = \sum_{i=1}^{n} \left(x_{1i} - \overline{x}_1 \right)^2$

Confidence interval for expected value at x_0 :

$$\hat{y}_0 \pm t_{\frac{\alpha}{2},n-2} s \sqrt{\frac{1}{n} + \frac{\left(x_0 - \overline{x}_1 \right)^2}{S_{xx}}}$$

Prediction interval at x_0

$$\hat{y}_0 \pm t_{\frac{\alpha}{2},n-2} s \sqrt{1 + \frac{1}{n} + \frac{\left(x_0 - \overline{x}_1 \right)^2}{S_{xx}}}$$

Multiple Linear regression

$$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

$$\varepsilon_i \sim N\left(0, \sigma^2\right) \text{ and independent}, i = 1, 2 \ldots, n$$

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

$$Cov\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ and fitted model: $\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{b}$

$$\hat{\sigma}^2 = S^2 = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-k-1}$$

Partitioning of variation

$$\sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2 = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 + \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}_i\right)^2$$

or $SS_T = SS_E + SS_R$

Inference Multiple regression

Confidence interval for the mean in x_0, Multiple regression

$$\hat{y}_0 \pm t_{\frac{\alpha}{2}, n-k-1}\, s\sqrt{\boldsymbol{x}_0'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0}$$

Prediction interval for a new observation in x_0, Multiple regression

$$\hat{y}_0 \pm t_{\frac{\alpha}{2},n-k-1} s\sqrt{1+\boldsymbol{x}_0^{'}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0}$$

Partial F-tests

$H_0 : \beta_{i+1} = \beta_{i+2} =,\ldots,= \beta_k = 0$ H_1 : at least one is different from zero.

$$F = \frac{\frac{SS_R\left(\beta_1,\ldots,\beta_k\right) - SS_R\left(\beta_1,\ldots,\beta_i\right)}{k-i}}{\frac{SS_E\left(\beta_1,\ldots,\beta_k\right)}{n-k-1}} \sim F_{k-i,n-k-1}$$

Forward selection

$$\max F_{in} = \frac{\frac{SS_R\left(\beta_1,\ldots,\beta_{i+1}\right) - SS_R\left(\beta_1,\ldots,\beta_i\right)}{1}}{\frac{SS_E\left(\beta_1,\ldots,\beta_{i+1}\right)}{n-(i+2)}} \sim F_{1,n-(i+2)}$$

Backward elimination

$$\min F_{in} = \frac{\frac{SS_R\left(\beta_1,\ldots,\beta_{i+1}\right) - SS_R\left(\beta_1,\ldots,\beta_i\right)}{1}}{\frac{SS_E\left(\beta_1,\ldots,\beta_{i+1}\right)}{n-(i+2)}} \sim F_{1,n-(i+2)}$$

Residuals in multiple linear regression

$\hat{\boldsymbol{\varepsilon}} = \left(\boldsymbol{I} - \boldsymbol{H} \right) \boldsymbol{\varepsilon}$ where $\boldsymbol{H} = \boldsymbol{X} \left(\boldsymbol{X}'\boldsymbol{X} \right)^{-1} \boldsymbol{X}'$

Studentized residuals:

$$r_i = \frac{e_i}{s\sqrt{1-h_{ii}}}$$

Check of

1. Outliers
2. Heterogenity in variance
3. Misspecified model
4. Normal distribution

Quality of fit

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2}$$

$$R^2_{adj} = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \cdot \frac{n-1}{n-k-1}$$

Quality of prediction

$$R^2_{pred} = 1 - \frac{\sum_{i=n}^{n}\frac{\left(y_i - \hat{y}_i\right)^2}{\left(1-h_{ii}\right)^2}}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \cdot$$

Orthogonal design matrix

X orthogonal

$$\hat{\beta}_i = \left(\boldsymbol{x}_i^{'}\boldsymbol{x}_i\right)^{-1}\boldsymbol{x}_i^{'}\boldsymbol{Y}$$

$$SS_R\left(\beta_0,\ldots,\beta_k\right) = \sum_{i=1}^{k} R\left(\beta_i\right)$$

Two level designs

Forsøksnr.	Temperatur	Mengde karbon i %	Prosentvis fjærer utan sprekkdanning
1	1450 F	0.5	67
2	1600 F	0.5	79
3	1450 F	0.7	61

4	1600F	0.7	75

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i} x_{2i}$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1450 & 0.5 & 725 \\ 1 & 1600 & 0.5 & 1015 \\ 1 & 1450 & 0.7 & 1120 \\ 1 & 1600 & 0.7 & 7200 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} 67 \\ 79 \\ 61 \\ 75 \end{bmatrix}, \quad \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_{12} \end{bmatrix} = \begin{bmatrix} 14.3 \\ 0.047 \\ -126.7 \\ 0.067 \end{bmatrix}$$

$$x_A = \frac{A - 1525}{75}, \; x_B = \frac{B - 0.6}{0.1}$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \quad \begin{bmatrix} b_0^c \\ b_1^c \\ b_2^c \\ b_{12}^c \end{bmatrix} = \begin{bmatrix} 78.5 \\ 6.5 \\ -2.5 \\ 0.5 \end{bmatrix}$$

2^3 experiment on standardform

A	B	C	AB	AC	BC	ABC	levelcode	y
-	-	-	+	+	+	-	1	60
+	-	-	-	-	+	+	a	72
-	+	-	-	+	-	+	b	54
+	+	-	+	-	-	-	ab	68
-	-	+	+	-	-	+	c	52
+	-	+	-	+	-	-	ac	83
-	+	+	-	-	+	-	bc	45
+	+	+	+	+	+	+	abc	80

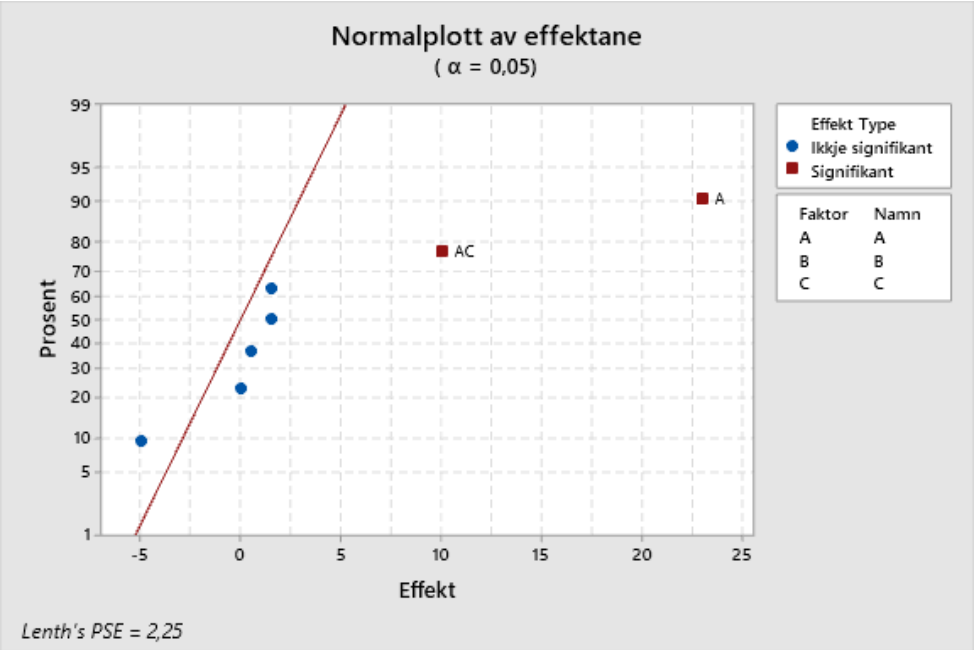
>Calulation of effects

$$\bar{y}_H - \bar{y}_L$$

Variance

$$\mathrm{Var}(\bar{y}_H - \bar{y}_L) = \frac{4\sigma^2}{n}$$

Evaluation of significance

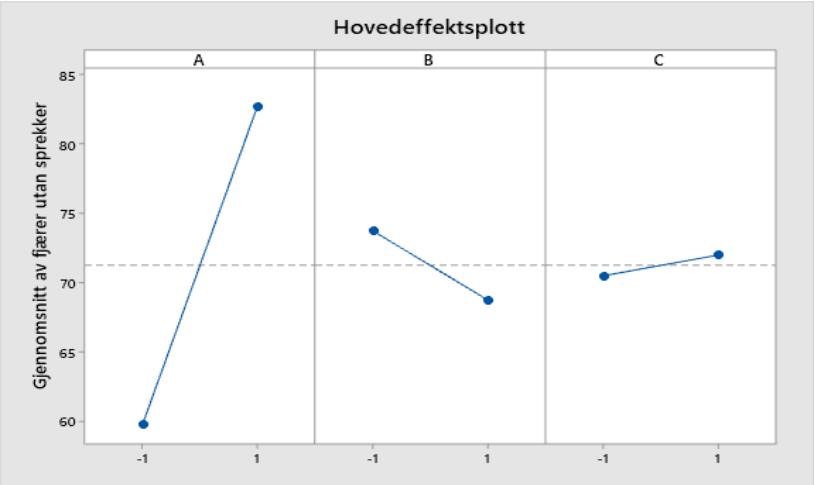


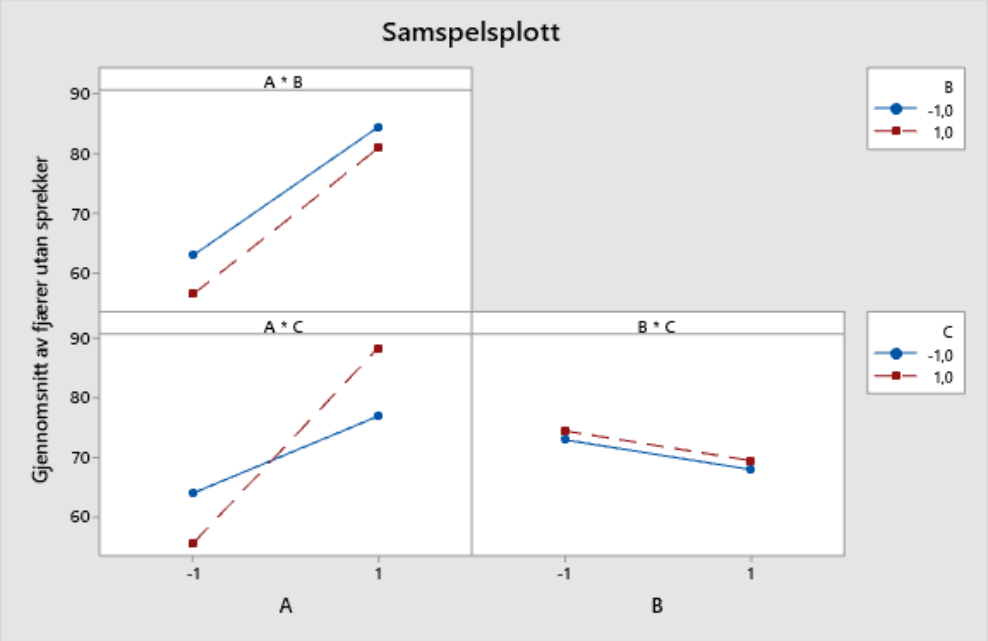
σ^2_{effekt} **is given by:** $\dfrac{\mathrm{AB\hat{C}}^2 + \mathrm{AB\hat{D}}^2 + \mathrm{AC\hat{D}}^2 + \mathrm{BC\hat{D}}^2 + \mathrm{ABC\hat{D}}^2}{5}$

Replicates

Average over $\sum_{j=1}^{m} \dfrac{(Y_{ij} - \bar{Y}_i)^2}{m-1}, i = 1, 2, \ldots, n$

Interpretation





Blocking and Fractioning

Define Block generators:

8 Blocks: $B_1 = ACE$ $B_2 = ABEF$ and $B_3 = ABCD$.

$$B_1B_2 = ACEABEF = BCF$$
$$B_1B_3 = ACEABCD = BDE$$
$$B_2B_3 = ABEFABCD = CDEF$$
$$B_1B_2B_3 = ACEABEFABCD = ADF$$

Fractional factorials

Find defining relation from generators.

2^{5-2}: D=AB, E=AC

I=ABD and I=ACE and I^2 =I=ABDACE=BCDE.

Hence the defining relation is I=ABD=ACE=BCDE.

2^{7-4}: D=AB, E=AC, F=BC and G=ABC

$$I = ABD=ACE=BCF=ABCG$$

I^2 =I=BCDE=ACDF=CDG=ABEF=BEG=AFG

I^3 =I=DEF=ADEG=BDFG=CEFG

I^4 =I=ABCDEFG

Analysis of variance

One-way

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \begin{cases} N\left(0, \sigma^2\right), \text{ and independent} \\ i = 1,2,\ldots,\text{k}, j = 1,2,\ldots,n_i \end{cases}$$

Hypothesis

$$H_0 = \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0 \quad H_1: \text{at least one } \alpha_i \neq 0.$$

Source	SS	DF	MS	F
Treatment	$SS_A = \sum_{i=1}^{k} n_i \left(\bar{y}_{i.} - \bar{y}_{..}\right)^2$	k-1	$SS_A / \ k$-1	$\frac{SS_A}{k\text{-}1} \Big/ \frac{SS_E}{n\text{-}k}$
Error	$SS_E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i.}\right)^2$	n-k	$SS_E / \ n\text{-}k$	
Total	$SS_T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{..}\right)^2$	n-1		

Randomized complete block design

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \begin{cases} N\ \ 0, \sigma^2 \quad \text{and independent} \\ j = 1,2,\cdots,b, \ \ i = 1,2,\cdots,k \end{cases}$$

Hypothesis

$$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0 \quad H_1: \text{at least two are different from zero}$$

Sources	SS	DF	MS	F
Treatment	$SS_A = b \sum_{i=1}^{k} \left(\bar{y}_{i.} - \bar{y}_{..}\right)^2$	k-1	$SS_A / \ k-1$	$\frac{SS_A}{k-1} \Big/ \frac{SS_E}{b-1 \quad k-1}$
Block	$SS_B = k \sum_{j=1}^{b} \left(\bar{y}_{.j} - \bar{y}_{..}\right)^2$	b-1	$SS_B / \ b-1$	
Error	$SS_E = \sum_{i=1}^{k} \sum_{j=1}^{b} \left(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}\right)^2$	(b-1)(k-1)	$SS_E / \ b-1 \quad k-1$	
Total	$SS_T = \sum_{i=1}^{k} \sum_{j=1}^{b} \left(y_{ij} - \bar{y}_{..}\right)^2$	bk-1		

Two-way analysis of variance

Model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \ \alpha\beta\ _{ij} + \varepsilon_{ijk} \begin{cases} \mathbf{N}\ 0,\sigma^2 \quad \text{and independent} \\ i = 1,2,\cdots,a,\ \ j = 1,2,\cdots,b,\ \ k = 1,2,\cdots,n \end{cases}$$

1. $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0 \quad H_1$: at least one is different from 0.
2. $H_0 : \beta_1 = \beta_2 = \cdots = \beta_b = 0 \quad H_1$: at least one is different from 0.
3. $H_0 : \alpha\beta_{11} = \alpha\beta_{12} = \cdots = \alpha\beta_{ab} = 0 \quad H_1$: at least one is different from 0.

Sources	SS	DF	MS	F
A	$SS_A = nb\sum_{i=1}^{a}\ y_{i..} - y_{...}\ ^2$	$a-1$	$\frac{SS_A}{a-1}$	$F = \frac{MS_A}{MS_E}$
B	$SS_B = na\sum_{j=1}^{b}\ y_{.j.} - \bar{y}_{...}\ ^2$	$b-1$	$\frac{SS_B}{b-1}$	$F = \frac{MS_B}{MS_E}$
Interaction AB	$SS_{AB} = n\sum_{i=1}^{a}\sum_{j=1}^{b}\ y_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}\ ^2$	$a-1\ \ b-1$	$\frac{SS_{AB}}{a-1\ \ b-1}$	$F = \frac{MS_{AB}}{MS_E}$
Error	$SS_E = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\ y_{ijk} - \bar{y}_{ij.}\ ^2$	$ab\ \ n-1$	$\frac{SS_E}{ab\ \ n-1}$	
Total	$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\ y_{ijk} - \bar{y}_{...}\ ^2$	$abn-1$		

Control Charts

Let T be a test statistics based on a sample $X_1, X_2, \ldots, X_n$.

Control Chart: LCL = $\mu_T - k\sigma_T$, CL = μ_T, UCL = $\mu_T + k\sigma_T$

Reduce the probability of false alarm: increase k.

Increase the probability of detecting out of control: increase n.

$\bar{X}bar - R$ C*hart*

k samples: Test statistics: $\bar{X}_1, \ldots, \bar{X}_k$, $R_1, \ldots, R_k$.

$$\bar{\bar{X}} = \frac{\sum_{i=1}^{k} \bar{X}_i}{k}, \qquad \bar{R} = \frac{\sum_{i=1}^{k} R_i}{k}.$$

$\bar{X}bar$: LCL= $\bar{\bar{X}} + \frac{3\bar{R}}{d_2\sqrt{n}}$, CL = $\bar{\bar{X}}$, UCL= $\bar{\bar{X}} + \frac{3\bar{R}}{d_2\sqrt{n}}$.

R C*hart*: LCL = $\bar{R} - \frac{3\bar{R}d_3}{d_2}$, CL = $\bar{R}$, UCL = $\bar{R} + \frac{3\bar{R}d_3}{d_2}$

p-charts

LCL= $\bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$, CL = $\bar{p}$, UCL= $\bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$

C-Charts

LCL= $\hat{\lambda} - 3\sqrt{\hat{\lambda}}$, CL = $\hat{\lambda}$, UCL = $\hat{\lambda} + 3\sqrt{\hat{\lambda}}$

χ^2 - tests

GOF

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(O_i - e_i\right)^2}{e_i} \approx \chi^2\left(k-1\right)$$

$e_i = np_i$

$r \times c$ Contingency Tables

Test for independence : $H_0 : p_{ij} = p_i.p_j \ \forall i, j$

Test for equal proportions in populations:
$H_0 : p_{ij} = p_i \ \forall j$ and for all i.

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(O_{ij} - e_{ij}\right)^2}{e_{ij}} \approx \chi^2\left(\left(r-1\right)\left(c-1\right)\right)$$

$$e_{ij} = \frac{n_i.n_j}{n}$$

Nonparametric tests

Sign test (continuous distribution)

X = Number of + (or -) for $x_i - \tilde{\mu}_0,\ i$=1,2,⋯,n

Test statistics: $X \sim B(n, 0.5)$

Wilcoxon 1 sample test (continuous and symmetric distribution)

Rank $|x_i - \tilde{\mu}_0|,\ i$=1,2,⋯,n in ascending order with ranks given by $1, 2, \cdots, n$.

Test statistics: $W_+ = \sum_{i: x_i - \tilde{\mu}_0 > 0} R_i$ and $W_- = \sum_{i: x_i - \tilde{\mu}_0 < 0} R_i$

Wilcoxon 2 sample test (continuous and identical distributions except for location)

Rank all observations in ascending order, in total $n_1 + n_2$

Test statistics: $U_1 = \sum_{x_i \in sample\ 1} R_i - \frac{n_1(n_1+1)}{2}$ and

$$U_2 = \sum_{x_i \in sample\ 2} R_i - \frac{n_2(n_2+1)}{2}$$

Kruskal Wallis test for number of samples, k >2 (continuous and identical distributions except for location)

Rank all observations in ascending order, in total $n_1 + n_2 + \ldots + n_k = n$

Compute average rank, R_i, for each sample.

Test statistics: $H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) \approx \chi^2(k-1)$